processes

MDPI

Article

Evaluation Techniques for Shale Oil Lithology and Mineral Composition Based on Principal Component Analysis Optimized Clustering Algorithm

Wenyuan Cai [1,2], Rui Deng [1,*], Chengquan Gao [3], Yingjie Wang [1], Weidong Ning [1,2], Boyu Shu [1] and Zhanglong Chen [2]

[1] CNPC Key Laboratory of Well Logging, Yangtze University, Wuhan 430100, China
[2] CNPC Logging Co., Ltd., Xi'an 710077, China
[3] PetroChina Tuha Oilfield Company, Hami 839009, China
* Correspondence: dengrui@yangtzeu.edu.cn; Tel.: +86-18-6278-07056

Abstract: Shale oil reservoirs are characterized by complex lithology, complex mineral composition and strong heterogeneity. This causes great difficulty in lithologic evaluation. In this paper, a method of lithology identification is proposed by means of intersection plot method and machine learning method, and lithology evaluation is carried out by combining the calculation of mineral content with a multi-mineral optimization model. The logging response characteristics of five lithologies are analyzed by using the logging curves selected by principal component analysis (PCA) discriminant analysis. In lithology identification, the system clustering algorithm is selected to identify shale oil reservoir lithology through layer-by-layer subdivision of sample lithology classification. Logging data has high vertical resolution and good continuity, and mineral prediction using logging data can ensure high accuracy. In this paper, the method of calculating mineral content by using multi-mineral optimization model has achieved good results in practice.

Keywords: lithological evaluation; principal component analysis; systematic clustering method; optimal multi-mineral model

Citation: Cai, W.; Deng, R.; Gao, C.; Wang, Y.; Ning, W.; Shu, B.; Chen, Z. Evaluation Techniques for Shale Oil Lithology and Mineral Composition Based on Principal Component Analysis Optimized Clustering Algorithm. *Processes* **2023**, *11*, 958. https://doi.org/10.3390/pr11030958

Academic Editors: Kejun Wu and An Su

Received: 15 February 2023
Revised: 12 March 2023
Accepted: 14 March 2023
Published: 21 March 2023

1. Introduction

Shale oil reservoir exploration and development started relatively late in China, and the progress is relatively slow. Although some breakthroughs have been made, a comprehensive shale oil reservoir evaluation method system has not been formed in general [1,2]. At present, the shale oil reservoir evaluation system that has been formed mainly refers to the logging evaluation methods of shale gas, tight oil and gas and other unconventional oil and gas. As exploration and development put forward deeper requirements for logging technology, it has begun to upgrade from "four characteristics" to "seven characteristics" evaluation, and on this basis, the "three quality" comprehensive evaluation of the reservoir was carried out [2,3]. Compared with tight oil, shale oil reservoirs are more complex, with higher requirements for evaluation of mobility and compressibility and more parameters.

Lithology evaluation, as an important part of this, includes lithology identification and rock mineral composition calculation. The lithologic identification method requires strong regional experience, and needs to be established by combining specific lithologic categories and logging response features to extract characteristic parameters that can distinguish the main lithologic categories. At present, lithology identification is mainly carried out through the crossplot method and machine learning methods, such as gradient lifting decision tree (GBDT) algorithm, PSO-SVM method, convolution-based cyclic neural network and ensemble learning, LSTM cyclic neural network, and BP neural network model optimization using principal component analysis.

Shale oil reservoir lithology is complex and changeable, its mineral composition is diverse, and its formation heterogeneity is strong. Therefore, its lithology evaluation is extremely difficult [3,4]; a set of evaluation methods for lithology identification has not yet been formed in the study area. Moreover, the calculation model of shale content and mineral composition, as well as the fracturing index characterizing the reservoir, have not been established [5].

Lithology evaluation is a very important basic research work in shale oil reservoir logging evaluation. Its results can provide basic laws and cognitive support for establishing the logging calculation method of key parameters of shale oil. Fine lithologic evaluation serves as the basis for the evaluation of reservoir physical properties and oil-bearing properties, and can also provide a reliable basis for the fracturing and later development of the reservoir. In terms of lithology evaluation, its main contents are lithology classification and naming, and determination of mineral components. The corresponding qualitative identification of lithology, quantitative calculation of mineral components and accurate calculation of brittleness index are particularly critical [6,7].

2. Lithology Determination and Determination of the Mineral Fraction of the Formation

2.1. Determination of Reservoir Lithology

On the basis of previous research, various factors were carefully analyzed and summarized. Comprehensive geological data, core analysis data, etc., combined with 448 thin section data points in the area, were analyzed and sorted. Then, the shale oil lithofacies division scheme of layer C with logging operability was determined. For subsequent research, the lithology of this interval has been given a simplified name. Therefore, the lithology of layer C is divided into the following five categories: (1) medium natural gamma siliceous shale (fine sandstone), (2) high natural gamma siliceous shale (muddy siltstone), (3) high natural gamma argillaceous shale (dark mudstone), (4) ultra-high natural gamma siliceous shale (black shale), and (5) high natural gamma tuff shale (tuff). It can be seen from Figure 1 that in the stratum, medium natural gamma siliceous shale (fine sandstone) and high natural gamma siliceous shale (muddy siltstone) account for the largest proportion. According to Figure 2, not all wells contain all five lithologies in layer C [5].

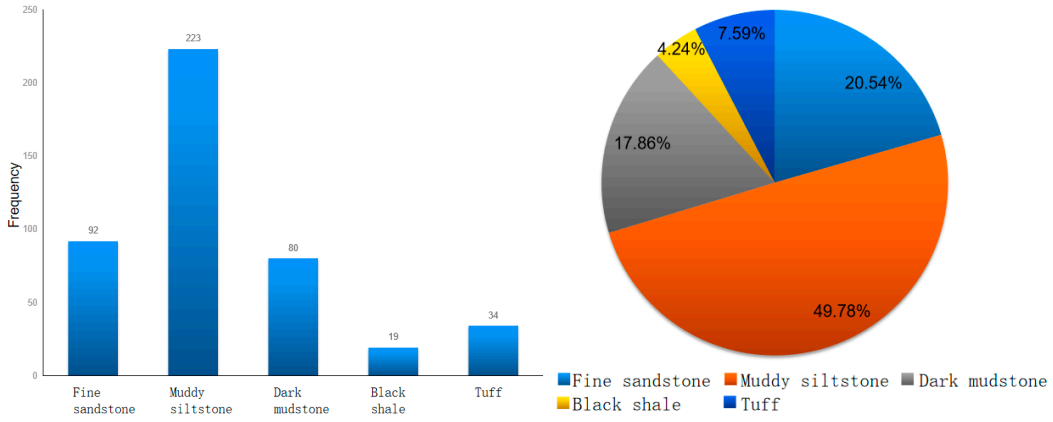

Figure 1. Statistical histogram and pie chart of the number of slices with different lithologies. (**a**) Histogram of the number of lithologic slices; (**b**) Statistical pie chart of lithology slice quantity.

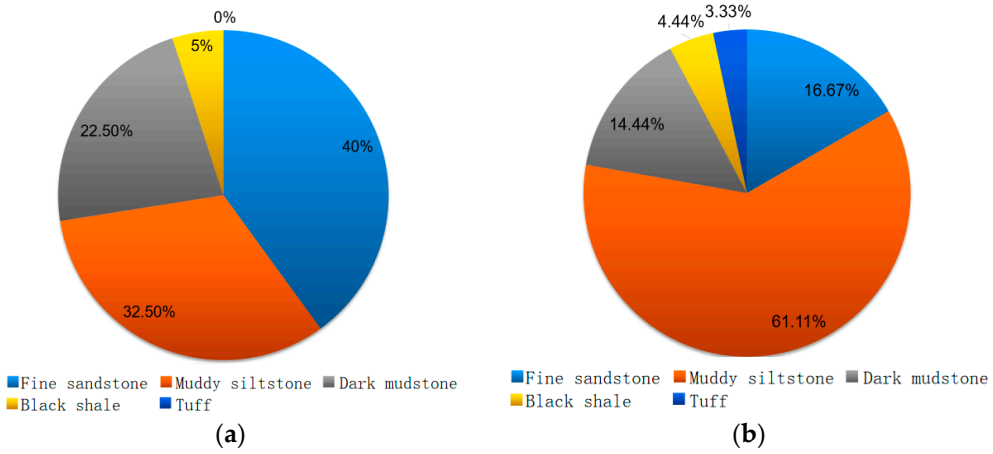

Figure 2. The proportion of different lithologic rocks in the stratum of Well J1 and Well J2. (**a**) Well J1; (**b**) Well J2.

2.2. Determination of Mineral Fractions

Argon ion polishing SEM experiments and X-ray crystal diffraction (XRD) experiments were carried out on a total of 30 samples from the C formation, including semi-quantitative XRD analysis of whole rock samples and semi-quantitative XRD analysis of clay minerals. Based on the results of the XRD and argon ion polishing SEM experiments (Figure 3), the mineral fractions and clay types of the C section can be determined. The results show that the mineral components in the study area are diverse, mainly including quartz, feldspar, clay minerals, calcite, carbonate particles and pyrite particles, as shown in Figure 3. The content percentage of each mineral and clay mineral component is shown in Figures 4 and 5, respectively. It is obvious that the main components of the mineral are quartz, clay mineral, plagioclase, potassium feldspar, pyrite and dolomite, and the average content is calculated to be 44.33%, 26.56%, 9.68%, 7.31%, 4.95% and 2.89%, respectively. The clay minerals are mainly composed of illite-montmorillonite mixed layer(I/S), illite, kaolinite and chlorite, and the average content is calculated to be 44.86%, 33.57%, 8.25% and 2.81%, respectively.

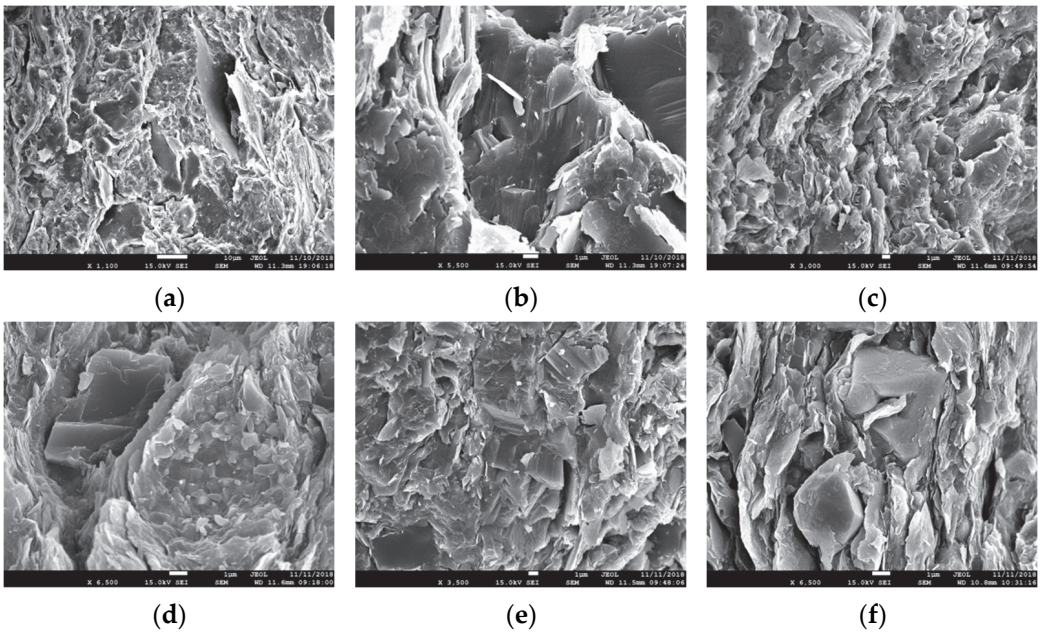

Figure 3. Scans of different minerals in shale oil formations. (**a**) Quartz grains; (**b**) Feldspar grains; (**c**) Clay minerals; (**d**) Calcite; (**e**) Carbonate particles; (**f**) Pyrite particles.

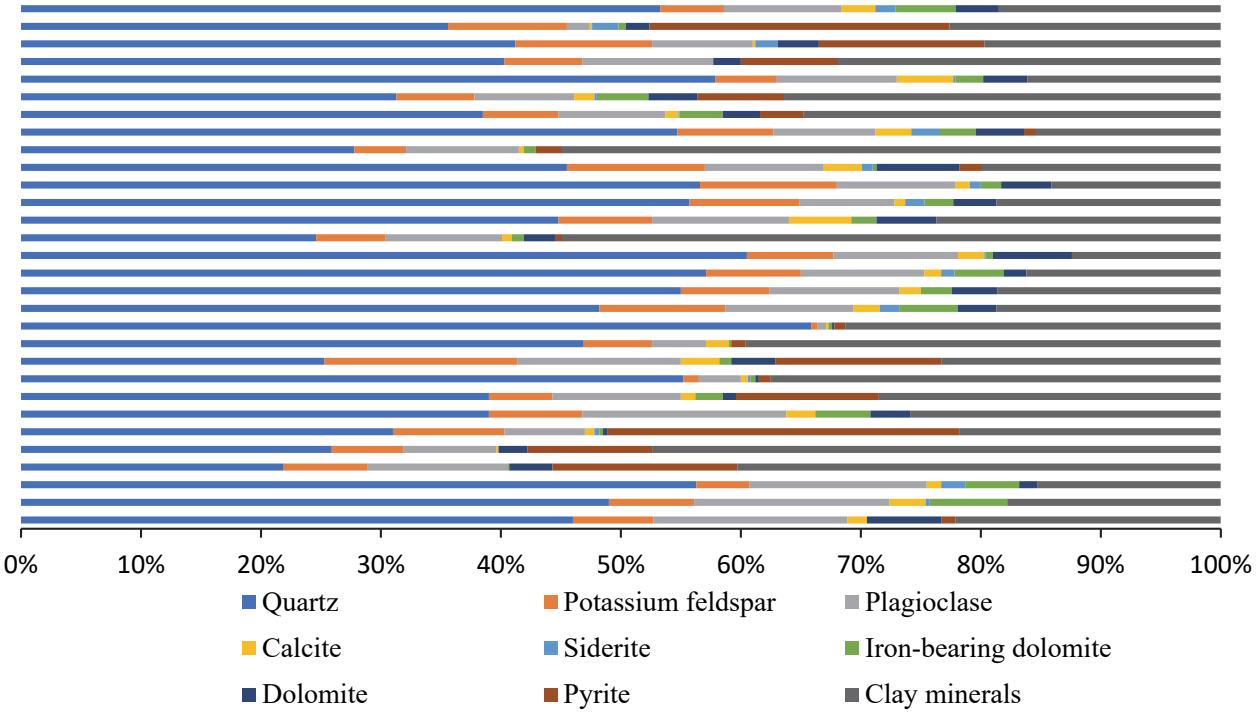

Figure 4. XRD mineral content percentage diagram of 30 rock samples in the study area.

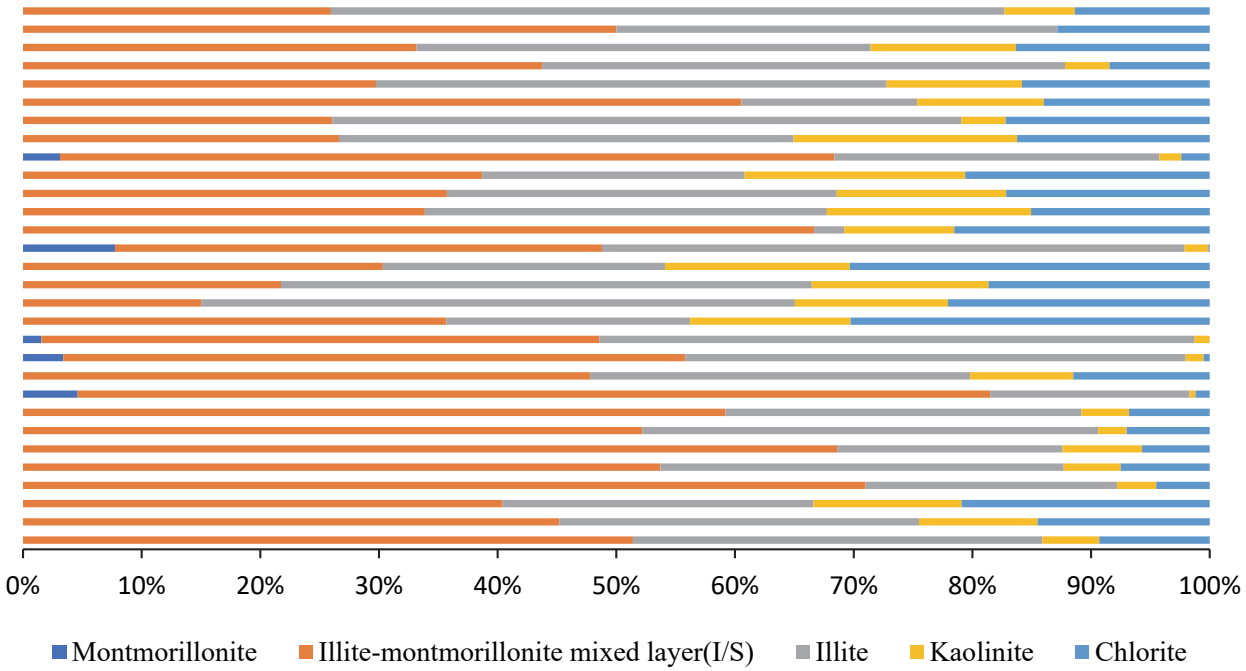

Figure 5. XRD chart of clay mineral content percentage of 30 rock samples in the study area.

3. Qualitative Lithology Identification Techniques Based on Principal Component Analysis Optimized Clustering Algorithms

The principal components analysis (PCA) technology was used to screen the logging curve, and the curve with a large correlation with the target curve was selected as the input curve of cluster analysis [8–12]. Ten logging curves including CNL, RT, RXO, DEN, GR, PE, AC, U, TH and K were input for PCA analysis. It can be seen that the correlation between TH curve and K curve is high, and the characteristics between them are not obvious. Although the correlation between Rt and Rxo is relatively low, Rxo is greatly affected by

cement sheath, borehole, mud filtrate, etc., and cannot fully reflect the characteristics of the actual reservoir. The independent characteristics between other curves are obvious, as shown in Figure 6. The PCA analysis results are shown in Table 1. The correlation between RT and other curves is significantly higher than that of RXO, and the correlation between TH curve and other curves is lower than that of K curve. The eight curves of RT, PE, CNL, DEN, GR, AC, K, U can be determined as lithologic cluster identification logging curves.

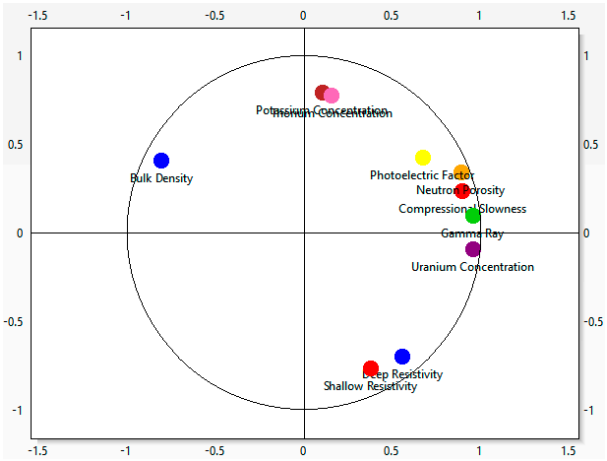

Figure 6. PCA analysis model of logging curve.

Table 1. Correlation analysis table of each logging curve.

	Bulk Density	Comprehension Slowness	Gamma Ray	Neutron Porosity	Photoelectric Factor	Deep Resistivity	Shallow Resistivity	Potassium Concentration	Thorium Concentration	Uranium Concentration
Bulk Density	1	−0.693173	−0.737776	−0.560614	−0.153228	−0.645156	−0.540367	0.274423	0.117068	−0.804864
Comprehension Slowness	−0.693173	1	0.805944	0.948787	0.610787	0.266069	0.0723262	0.21398	0.260662	0.772366
Gamma Ray	−0.737776	0.805944	1	0.8082	0.644885	0.429516	0.253071	0.140882	0.269512	0.972025
Neutron Porosity	−0.560614	0.948787	0.8082	1	0.761449	0.221603	0.0243947	0.300354	0.308405	0.759762
Photoelectric Factor	−0.153228	0.610787	0.644885	0.761449	1	0.11066	−0.0334943	0.402085	0.216005	0.598149
Deep Resistivity	−0.645156	0.266069	0.429516	0.221603	0.11066	1	0.89983	−0.314645	−0.336595	0.534918
Shallow Resistivity	−0.540367	0.0723262	0.253071	0.0243947	−0.0334943	0.89983	1	−0.320226	−0.398188	0.357767
Potassium Concentration	0.274423	0.21398	0.140882	0.300354	0.402085	−0.314645	−0.320226	1	0.703384	−0.0528857
Thorium Concentration	0.117068	0.260662	0.269512	0.308405	0.216005	−0.336595	−0.398188	0.703384	1	0.0649356
Uranium Concentration	−0.804864	0.772366	0.972025	0.759762	0.598149	0.534918	0.357767	−0.0528857	0.0649356	1

3.1. Optimal Selection of Sensitive Parameters Based on Principal Component Analysis Techniques

The logging curves were screened using principal components analysis (PCA) and those with a high correlation with the target curve were selected as input curves for the cluster [13–17]. A total of ten log curves, including CNL, RT, RXO, DEN, GR, PE, AC, U, TH and K, were inputted for PCA analysis, as shown in Figure 6 and Table 1.

According to the logging curves selected by PCA discriminant analysis, and verifying the correctness of PCA method, the logging response characteristics of five lithologies were analyzed by adding AF10 curve, as shown in Figure 7. The patterns of ultra-high natural gamma siliceous shale, high natural gamma argillaceous shale and high natural gamma tuffaceous shale are obviously different. The ultra-high natural gamma siliceous shale has the characteristics of high gamma, high resistance, low neutron and low potassium, as shown in Figure 7d. The clayey shale with high natural gamma has the characteristics of high potassium, high density and low gamma, as shown in Figure 7c. High natural gamma tuffaceous shale has the characteristics of high neutron, high density, low potassium and low resistance, as shown in Figure 7e. Medium natural gamma siliceous shale and

high natural gamma siliceous shale have similar logging response characteristics and poor discrimination effect, as shown in Figure 7a,b. When put in the shale formation for separate analysis, the results show that the natural gamma siliceous shale in the shale layer has the characteristics of relatively high resistance, low potassium and high neutron. High natural gamma siliceous shale has relatively high potassium, low resistance and low neutron in the shale layer, as shown in Figure 7f (where the red part of the figure is the high natural gamma siliceous shale and the blue part is the medium natural gamma siliceous shale). It can be seen that the AT10 curve does not respond significantly to the above lithological characteristics, which can verify the correctness of PCA method.

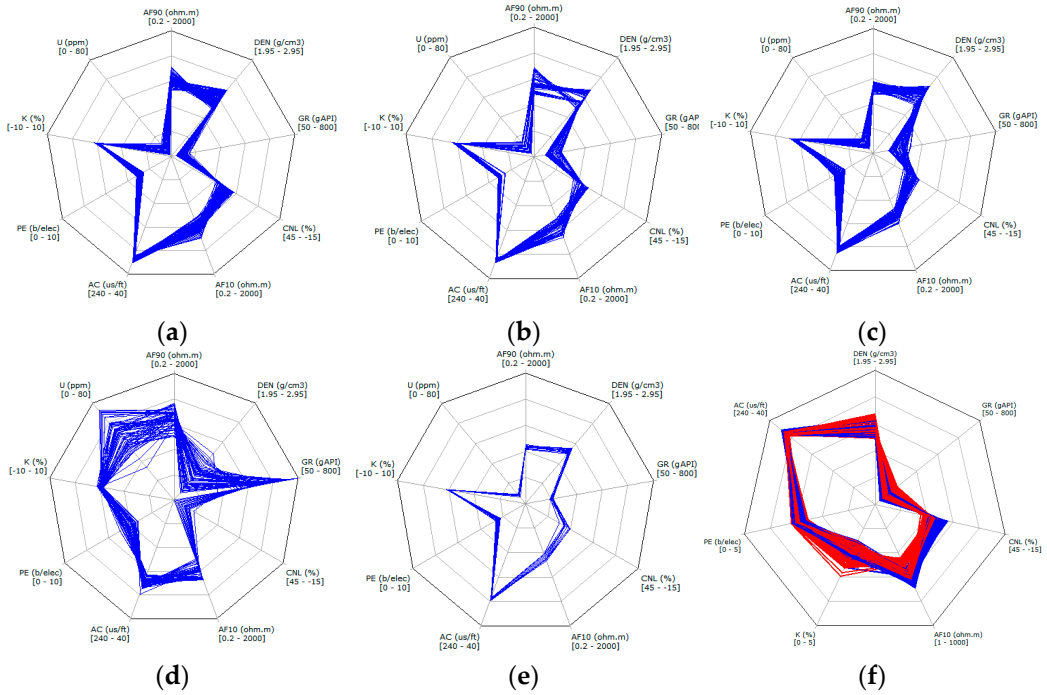

Figure 7. Radar chart of five kinds of lithology logging curve characteristic analysis. (**a**) Medium natural gamma siliceous shale; (**b**) High natural gamma siliceous shale; (**c**) High natural gamma clayey shale; (**d**) Very high natural gamma siliceous shale; (**e**) High natural gamma tuff shale; (**f**) Separate analysis of shale layers.

3.2. Lithology Identification Techniques for Logging Based on Cluster Analysis

Cluster analysis is an unsupervised classification algorithm that relies only on the similarity of things as the basis for classification. There are several methods such as systematic clustering, decomposition clustering, fuzzy clustering, dynamic clustering, legend clustering and clustering forecasting method. In this lithology identification, the systematic clustering method is chosen to subdivide the samples layer by layer in the process of lithology classification. The basic idea and principle are shown in Figure 8, where the input of logging parameters of samples with known lithology is firstly subjected to hierarchical clustering learning, then the lithology identification model is constructed, and finally the constructed model is applied to the logging data with unknown lithology for lithology identification [1,2,18–21]. The clustering analysis algorithm model and operation written in this paper are based on Python software.

First of all, the deep learning algorithm requires a large amount of data, so it needs rich lithological data as a sample learning well and as a cluster supervisor to complete the construction of the model and increase the recognition accuracy. Here, J3 well with rich core thin section data is taken as a sample for learning, and eight curve data corresponding to lithology of different depth layers are input into the algorithm model. As shown in Figure 9, it is the depth and curve data corresponding to the thin section lithology. The

legends of various lithologies in the log interpretation diagram are shown in Table 2. In order to use the lithologic identification model of J3 well to identify other wells, J3 well must be standardized as a standard well, so that the identification accuracy of the model can be improved. The lithologic identification model based on clustering algorithm to Well J2 is applied, so that it can complete the lithologic identification of Well J2 (Figure 10), and determine the accuracy of lithologic identification of unknown wells.

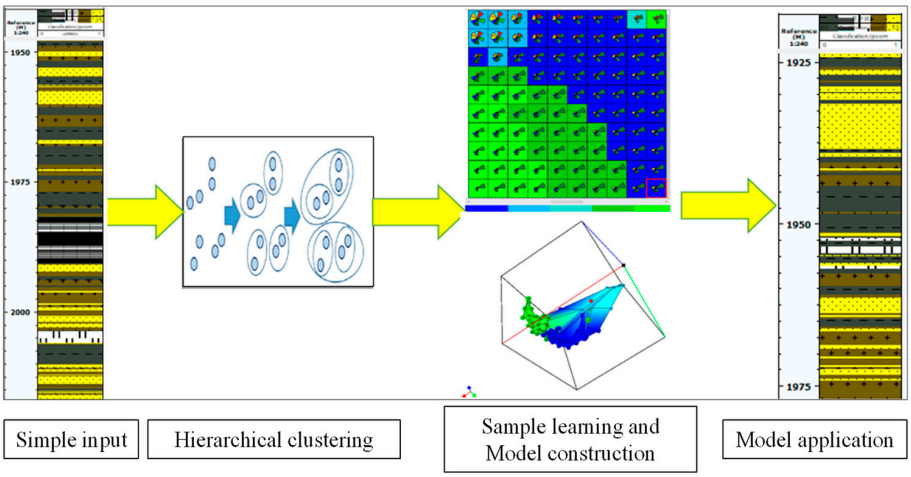

Figure 8. Flow chart of well logging lithology identification based on cluster analysis.

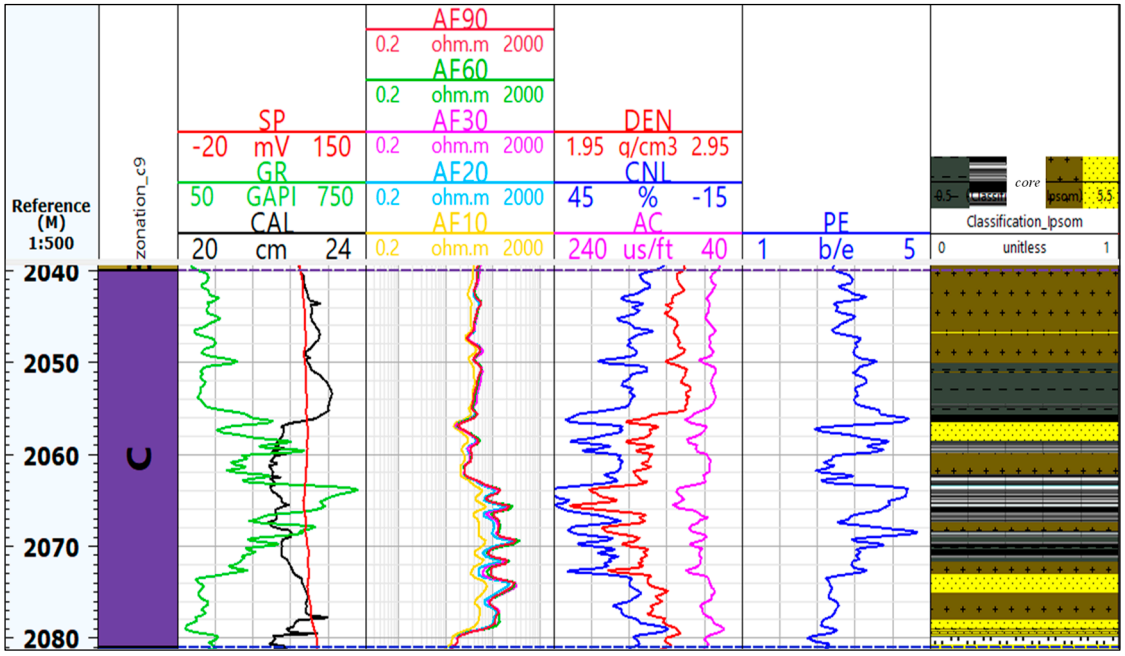

Figure 9. Results of lithology identification of well J3 in sample data well.

Table 2. Icon representation of different lithologies.

Medium Natural Gamma Siliceous Shale	High Natural Gamma Siliceous Shale	High Natural Gamma Clayey Shale	Extra High Natural Gamma Siliceous Shale	High Natural Gamma Tuffaceous Shale
				

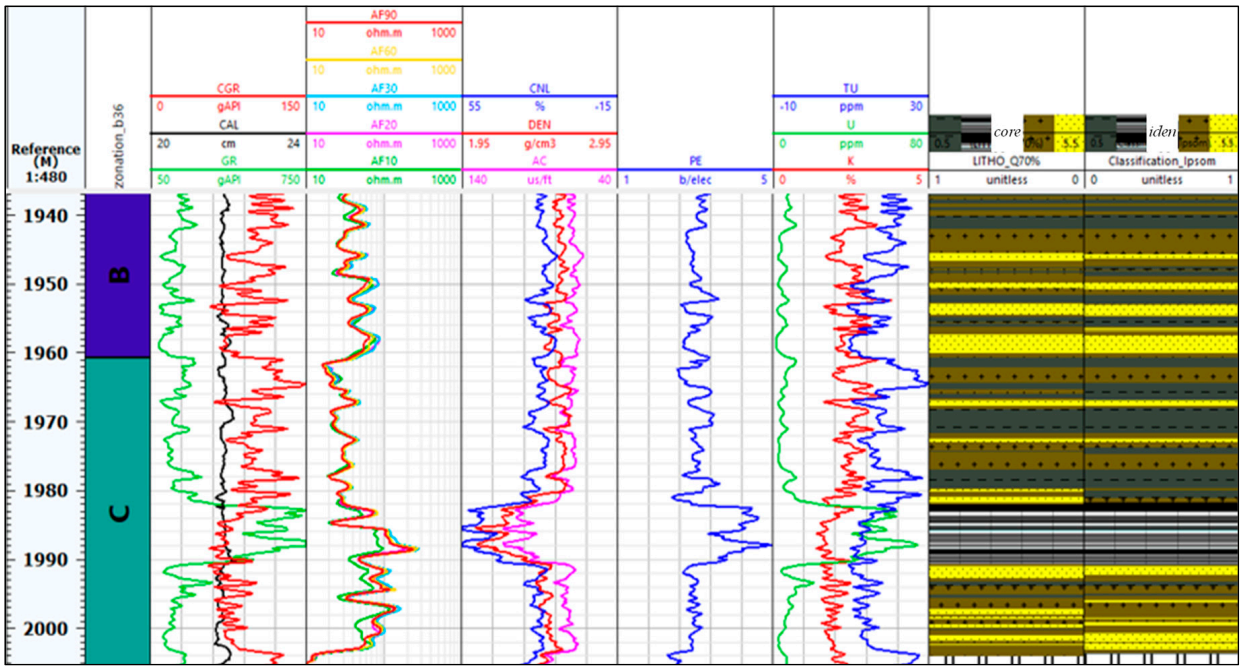

Figure 10. Log interpretation diagram for lithology identification in Well J2 of Applied Well.

This clustering algorithm is used to identify the lithology of different depth sections of Well J2, and 86.37% of the identified rock properties correspond to the actual core thin section lithology. There are 13.63% deep layer lithology identification errors, which is within the allowable range of industrial production. As shown in Figure 10, core is the thin section lithology, iden is the identification lithology. It can be considered that it is feasible to use J3 well as a sample and J2 well as a test set to verify the model in shale oil reservoir lithology identification. Well J3 can be used as a sample input and the algorithm can be extended to lithologic identification in the study area.

4. Quantitative Calculation of Mineral Composition

Calculation of mineral content is a complex and important work in the fine logging evaluation of reservoirs, and the evaluation results have important guiding significance for the evaluation of physical properties. The commonly used mineral content logging calculation methods mainly include the following methods: multi-mineral optimization calculation model, petrophysical modeling of a single mineral, lithologic capture logging, and element logging. However, when there are many kinds of minerals, it is difficult to calculate the mineral content using the petrophysical modeling method of a single mineral, and the cumulative error is too large. The element capture logging method is a new method with high accuracy to calculate mineral content, but it is mostly measured in key exploration wells. At present, the comprehensive popularization of all wells is still difficult in terms of funding and instrument quantity, so this method is not universal; the calculation of mineral content by element logging method has the same problem as that by element capture logging method. Its calculation accuracy is very high, but not all wells have element logging data. Formation C is rich in lithology, and a limited number of core mineral analysis cannot effectively describe the distribution characteristics of various mineral contents in shale. Therefore, on the basis of measured mineral content, more accurate classification is needed by other means. Logging data has high vertical resolution and good continuity, so the accuracy of mineral prediction using logging data is high. Compared with several methods, the method of calculating mineral content using a multi-mineral optimization model is a relatively good choice. The interpretation of the two well sections by the multi-mineral optimization model adopted in this paper depends on the TECHLOG software.

4.1. Calculation of Shale Content Based on Combination Method

Calculation of shale content is usually based on the radioactivity of the reservoir. Previous researchers have explored many methods for calculating shale content in conventional reservoirs, which can also ensure the accuracy and reliability of the results. This mainly includes the natural gamma method, natural gamma spectrum logging method and neutron-density intersection method. However, for shale oil reservoirs, the high content of rock debris and terrigenous debris, coupled with the development of highly radioactive minerals such as potassium feldspar in the C section of the formation, lead to inaccurate results of shale content calculation using GR parameters. Due to the complex and variable lithology in the C section formation, the same variety and complexity of mineral components, and the intercalation of thin layers, there are many influencing factors, so the accuracy of calculating shale content is poor if one of the four types of methods is chosen arbitrarily, and through previous experience, the calculation results of shale content are generally upper limits.

In shale development intervals, kerogen content and oil saturation are significantly higher than non-shale intervals. The formation will change under the influence of these factors, and these changes will be reflected in the response of logging curves. Using the different logging response characteristics of shale and non-shale reservoirs, it is basically possible to identify shale sections of sand-mud thin interbedded type by overlaying neutron and resistivity curves, density and neutron curves, overlaying neutron and PE curves, and envelope filling of PE and density curves (e.g., Figure 11). By dividing the shale and non-shale sections, and using the response characteristics of the logging curves as an entry point, we can avoid the interference of layering on the lithology calculation and choose a more targeted shale content calculation method according to the response characteristics.

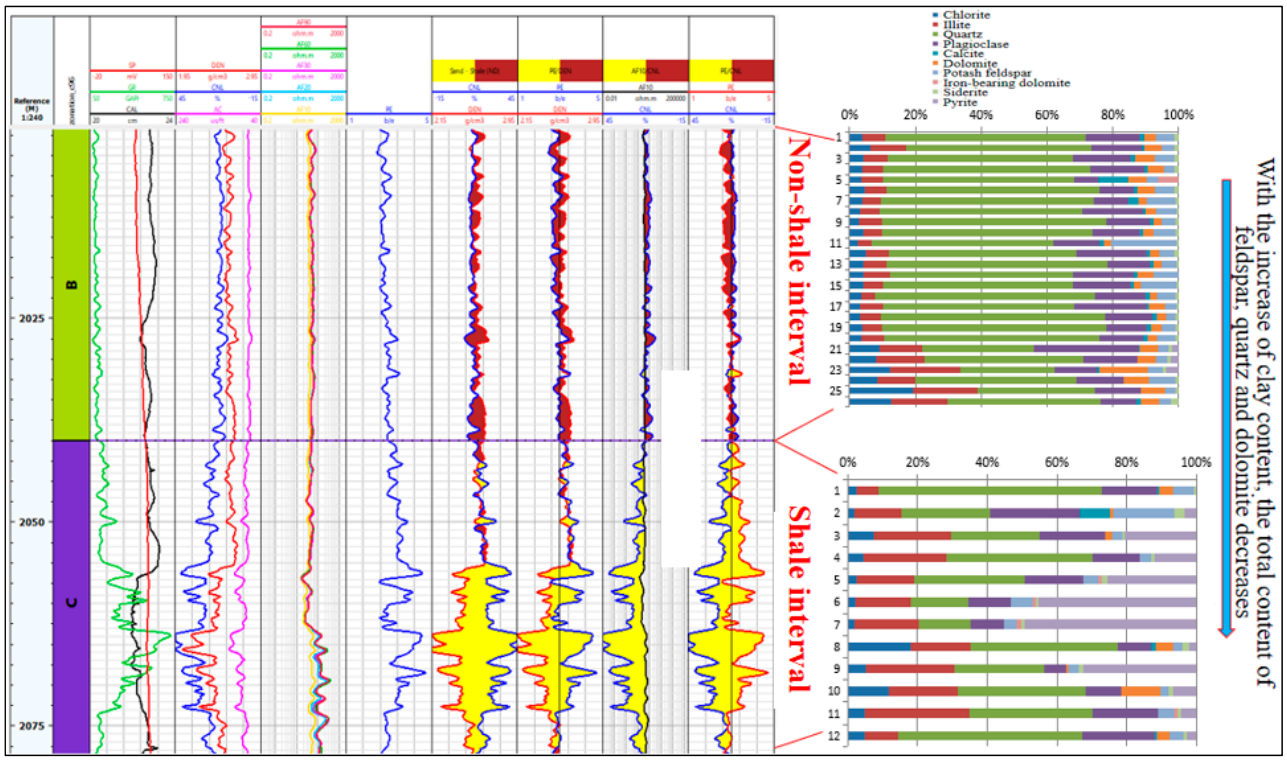

Figure 11. Well log curve and mineral content percentage diagram of Well J2.

Based on the division of shale interval and non-shale interval, the characteristics of logging curve are analyzed, the optimal calculation method of shale content in the formation is selected, and the accuracy of multi-mineral component calculation results is improved. According to the analysis, the clay content of the non-shale interval is low, as

shown in Figure 12. In the figure, the blue is the shale interval, and the yellow is the non-shale. In the non-shale section, the distribution range of neutron, density, acoustic transit time, natural gamma and PE values is small, and the calculation accuracy of conventional data curve is poor, so it is suitable to use the gamma curve to calculate shale content. The shale content in the shale interval increases, and the distribution range of neutron, density, acoustic transit time, natural gamma and PE values changes greatly, so it is suitable to calculate the shale content using the neutron-density combined gamma curve.

According to the regional experience, the effect of calculating shale content by removing uranium gamma curve (CGR) is better. The CGR curve has been measured in new wells, but for some old wells, the CGR curve of shale and non-shale layers has not been measured, so it is necessary to reconstruct the curve to calculate the shale content of this interval to make the calculation of shale content more accurate.

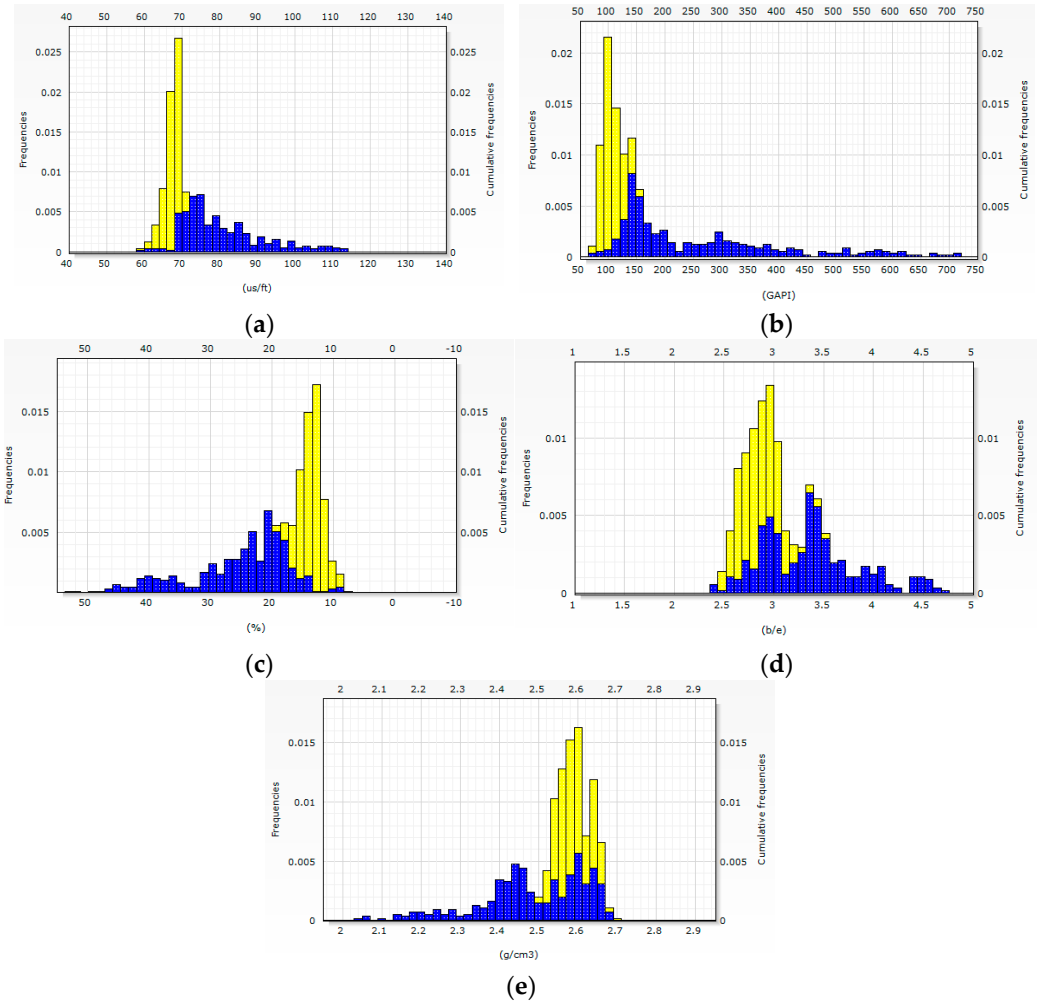

Figure 12. Different curve log response characteristics of shale and non-shale intervals. (**a**) Histogram of frequency distribution of AC curve; (**b**) Histogram of frequency distribution of GR curve; (**c**) Histogram of frequency distribution of CNL curve; (**d**) Histogram of frequency distribution of PE curve; (**e**) Histogram of frequency distribution of DEN curves.

In order to make the calculation result of shale content more accurate, shale oil formation identification should be carried out before calculation, and shale content of shale section is the average value of neutron density method and uranium removal gamma method. The shale content in the non-shale section is calculated using the uranium removal gamma curve to form the shale curve.

(1) Calculation of shale content by neutron-density crossplot

In conventional triple porosity logging curves, neutron and density logging are more sensitive to the logging response characteristics of changes in formation hydrocarbon flow and shale content than sonic time difference logging and are largely independent of the form of formation mud distribution. The neutron-density rendezvous method is therefore often used to calculate the shale content of high shale content formations, low pore low permeability formations and high natural gamma formations, and is calculated as follows.

$$V_{sh} = A/B \tag{1}$$

$$A = \rho_b(N_{ma} - 100) - CNL\cdot(\rho_{ma} - \rho_f) - N_{ma} \times \rho_f + \rho_{ma} \tag{2}$$

$$B = (\rho_{sh} - \rho_f)(N_{ma} - 100) - (N_{sh} - 100)(\rho_{ma} - \rho_f) \tag{3}$$

In the formula, N_{ma} is the neutron value of the rock skeleton, N_{sh} is the neutron value of the mudstone, CNL is the neutron value measured by the target layer section, %; ρ_{ma} is the density value of the rock skeleton, ρ_{sh} is the pure mudstone density value, ρ_f is the formation fluid density value, ρ_b is the density value measured by the target layer section, g/cm^3.

(2) Calculation of shale content by reconstructing uranium-free gamma curve method

The formation in the study area is rich in uranium, but for the high uranium formation, the conventional gamma curve cannot truly reflect the change of shale content in the formation. Usually, energy spectrum logging or element logging can be used to better identify such reservoirs. However, for the well section without energy spectrum and element logging, it is necessary to establish a quantitative relationship between the conventional logging information and the shale content of the formation. This curve is called the uranium-free gamma curve (CGR). Correlation analysis on CGR and GR, CNL, DEN, AC, AF90, PE and other logging curves (Table 3) should be carried out, and the multiple regression calculation model of log parameter reconstruction with good correlation selected.

Table 3. Correlation analysis of logging curves in interval c in the study area.

	AC	AF20	AF90	CNL	DEN	GR	PE	CGR
AC	1							
AF20	0.307215	1						
AF90	0.414528	0.96554	1					
CNL	0.952085	0.270417	0.384129	1				
DEN	−0.83244	−0.49978	−0.58967	−0.74327	1			
GR	0.910569	0.291836	0.396808	0.922276	−0.81952	1		
PE	0.664827	0.107605	0.17078	0.779276	−0.33254	0.713711	1	
CGR	0.309785	−0.3616	−0.33071	0.420375	0.009589	0.335664	0.414639	1

Based on the correlation analysis, the acoustic AC, neutron CNL, gamma GR and PE curves were selected to be fitted to the de-uranium gamma curve, and the multiple regression calculation model for the reconstruction of uranium-free gamma (CGR) curve (R^2 = 0.786) was

$$CGR = -0.58494AC + 3.42469CNL - 0.02413GR + 3.02460PE + 157.512$$

The formula for calculating the shale content using the reconstructed de-uranization gamma curve is

$$\Delta CGR = \frac{CGR - CGR_{\min}}{CGR_{\max} - CGR_{\min}} \tag{4}$$

$$V_{sh} = \frac{2^{GCUR \times \Delta CGR} - 1}{2^{GCUR} - 1} \tag{5}$$

In the formula, CGR represents the formation uranium-free gamma curve measurement value, CGR_{min} is the pure formation uranium-free gamma value, CGR_{max} is the pure mudstone uranium-free gamma value, Gapi and *GCUR* is the formation correction factor, which is 3.7 in the study area.

As shown in Figures 13 and 14, the results of the through-combination method of calculating shale content have small errors compared with experimental measurements and can meet the needs of fine calculation of shale content in the study area.

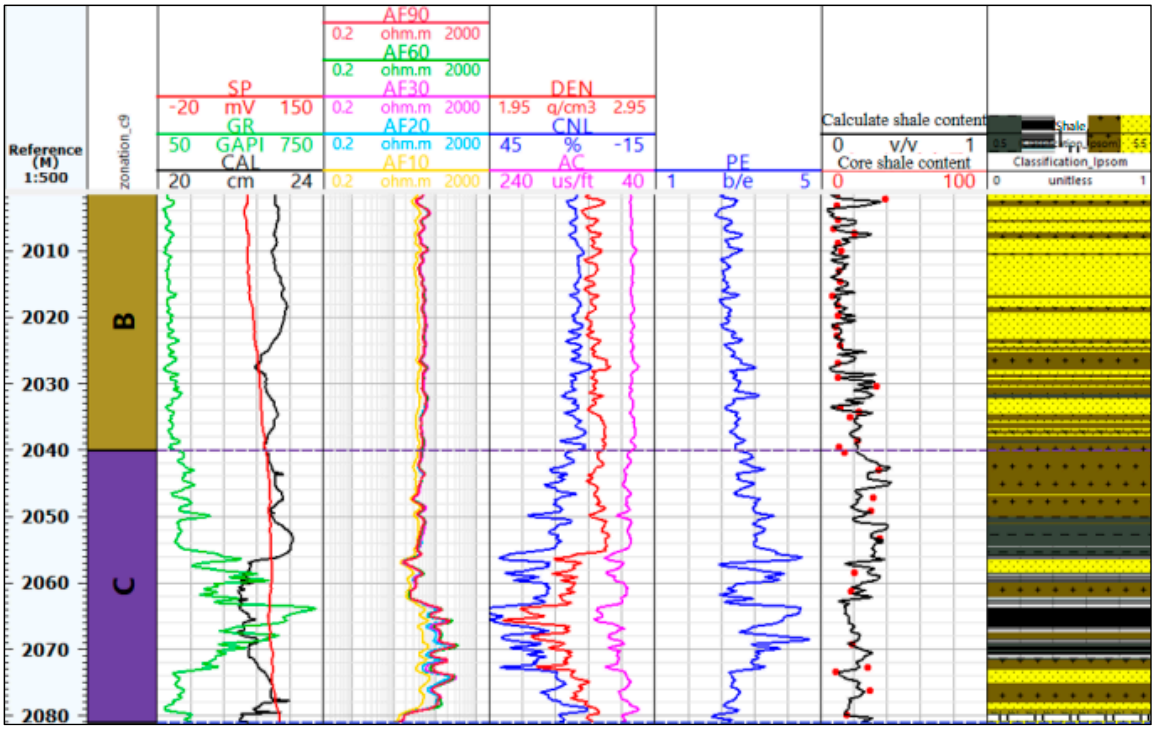

Figure 13. Log interpretation diagram of calculating shale content by combination method in Well J2.

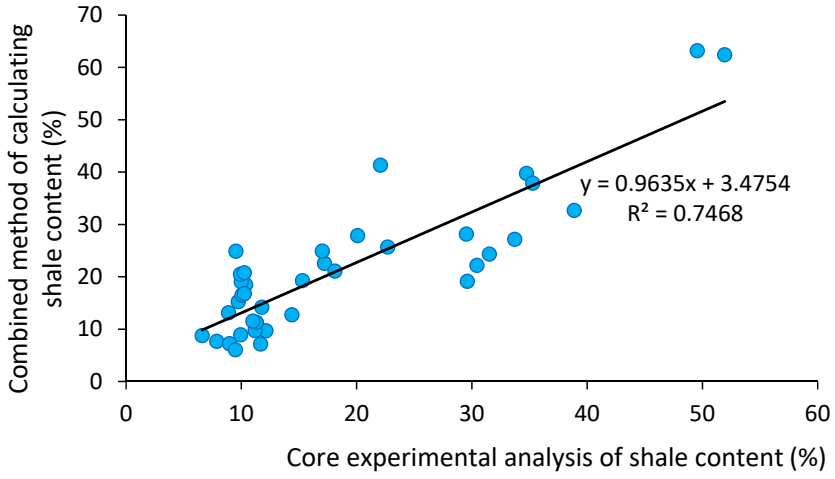

Figure 14. Cross diagram of shale content in core experiment analysis and shale content calculated by combination method.

4.2. Optimisation-Based Multi-Mineral Model Approach to Calculate Mineral Content

The petrophysical properties of the formation and its logging response mechanism are used to discover logging information and evaluate hydrocarbons. According to the petrophysical model of rock volume, the logging signal is derived from the rock skeleton and pore fluids, and its numerical magnitude is weighted and averaged according to its eigenvalues and the proportion of volume it occupies. Optimization of the multi-mineral model, based on the non-linear weighted least squares principle, establishes the target equation based on a suitable mineral volume model and reasonable reservoir logging parameters, and uses optimization techniques to continuously adjust the unknown reservoir parameter values by selecting initial values. Therefore, based on the above principles, the logging response equation general equation is established as

$$\log date = \sum_{i=1}^{n} v_i \times x_i \tag{6}$$

$$v_1 + v_2 + v_3 + \cdots + v_n = 1 \tag{7}$$

The logging response values of natural gamma, bulk density and acoustic time difference can be considered as the average of the physical quantities of the components of the response per unit volume of rock. Specific examples of logging response equations and objective functions for density, acoustic time difference and neutron are as follows.

$$\rho_b = \rho_1 v_1 + \rho_2 v_2 + \rho_3 v_3 + \cdots + \rho_n v_n \tag{8}$$

$$\Delta t = \Delta t_1 v_1 + \Delta t_2 v_2 + \Delta t_3 v_3 + \cdots + \Delta t_n v_n \tag{9}$$

$$CNL = CNL_1 v_1 + CNL_2 v_2 + CNL_3 v_3 + \cdots + CNL_n v_n \tag{10}$$

In the formula: $i = 1, 2, \ldots, n$, n is the number of mineral components and fluids in the formation; v_i is the percentage content of the i-th mineral to be determined, the sum of the volume percentage content of each component of the rock is 1, and they are all greater than or equal to 0; x_i is the logging response value of the i-th pure mineral; logdate is the measured logging curve of the formation, which is the comprehensive logging response value of all minerals, and its value is directly read from the logging curve. The percentages of each mineral can be found by combining ρ_i, Δt_i, and CNL_i are the log response values for density, acoustic and neutron for each mineral, respectively.

In theory, the number of solved minerals cannot be higher than the number of independent logging physical quantities to constrain the set of equations to have a higher accuracy in solving for minerals. Multi-mineral component calculations of lithology can be based on existing results by adding new parameter conditions to obtain more detailed and accurate calculations. The addition of porosity curves and water saturation curves to the model can make the calculation of the fluid component more accurate and the results of the in situ formation profile calculations more closely match the actual formation characteristics.

According to core data, thin section data and XRD data, it can be determined that the composition of rock minerals in the study area is eight skeleton minerals. These are quartz, feldspar, illite, chlorite, calcite, dolomite, pyrite and organic kerogen. Combined with the calculated shale content and the characteristics of the constituent minerals themselves, using the logging response equation comprehensively, under the given constraint conditions, the optimization algorithm can be used to obtain the approximate solution of mineral content. As shown in Figure 15, the results of the log interpretation of the mineral profile of well J3, which is in general agreement with the mineral content measured by XRD experiments. Figure 16 shows the log interpretation of the mineral content calculated by the optimized multi-mineral model of Well J2.

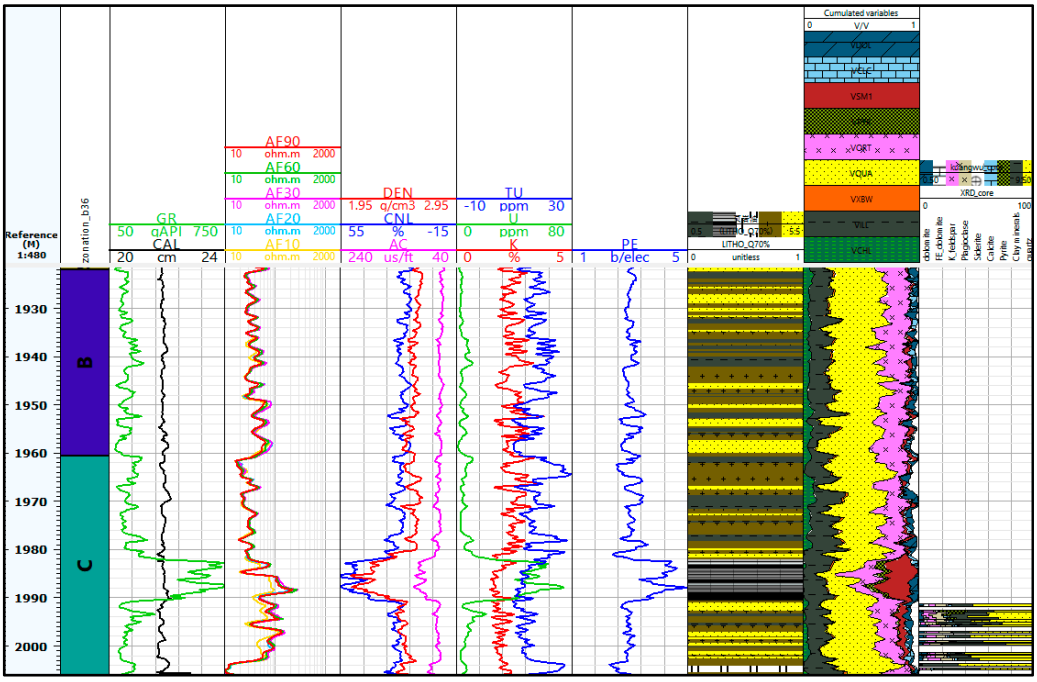

Figure 15. Log interpretation diagram of mineral content calculated by the optimized multi-mineral model in Well J3.

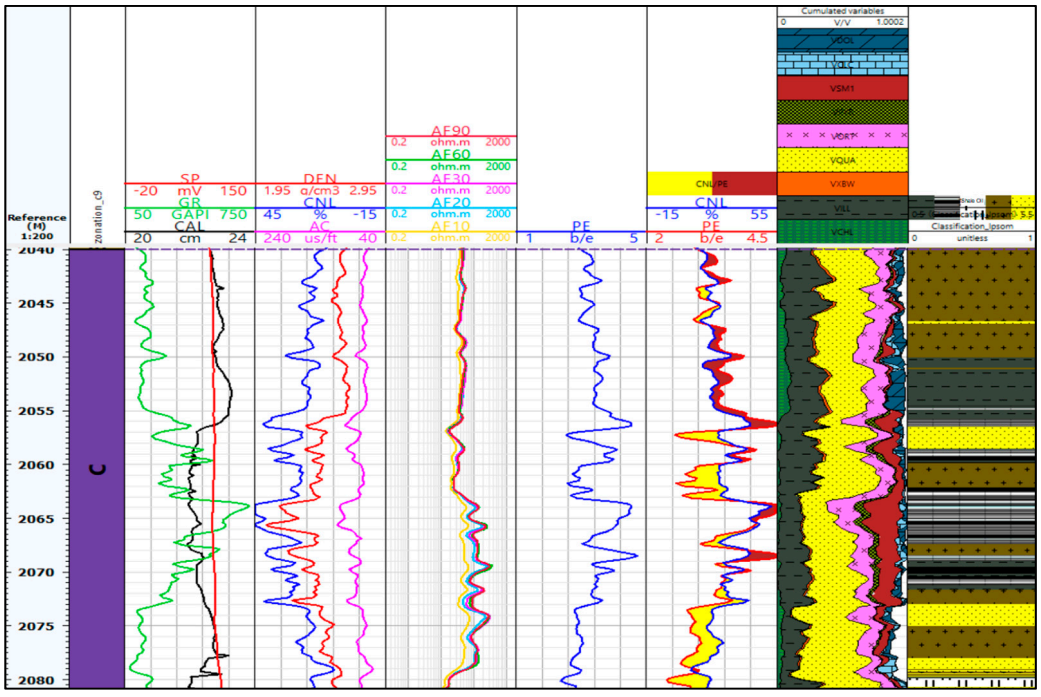

Figure 16. Log interpretation diagram of mineral content calculated by the optimized multi-mineral model in Well J2.

5. Conclusions

Given that reservoirs of different lithologies are often shown on one logging data as a combined effect of being shown on multiple logs, it is inevitable that statistical learning techniques will be used to explore lithology identification using multiple logging data.

Firstly, the principal component analysis algorithm is introduced to analyze the importance of the weight of each lithologic feature on the clustering analysis. After that, a shale oil reservoir lithology identification technology based on principal component

analysis and optimized clustering algorithm is established. Combined with the application of the optimized multi-mineral model, the quantitative calculation of the formation mineral components has been completed. This paper uses high-precision lithologic identification technology to lay a good foundation for the subsequent analysis of physical properties and oil-bearing properties and the establishment of the calculation model of key parameters of shale oil reservoirs.

In terms of the corresponding coincidence rate of core results, the lithologic identification based on principal component analysis has good effect and can be used in production practice.

Author Contributions: W.C., B.S. and C.G.: Conceptualization, Data curation, Methodology, Investigation, Validation, Writing—original draft. R.D.: Supervision, Formal analysis, Funding acquisition, Resources, Writing—review & editing. Y.W., W.N. and Z.C.: Supervision, Formal analysis, Resources. All authors have read and agreed to the published version of the manuscript.

Funding: The research is funded by the Major National Science and Technology Projects of China "Multidimensional and high precision imaging logging series" (No. 2017ZX05019001).

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author, RD.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Pan, T.; Ma, X.; Xie, A.; Gao, Z. Optimization of BP neural network model in sand and conglomerate lithology identification using principal component analysis. *Xinjiang Geol.* **2020**, *38*, 417–420.
2. Kong, Q.; Yang, C.; Li, H.; Geng, C.; Deng, J. A lithology identification method based on graph-theoretic clustering and minimum proximity algorithm—An example of carbonate reservoirs of the Leikoupo Formation in the western Sichuan Basin. *Pet. Nat. Gas Geol.* **2020**, *41*, 884–890.
3. Fang, Y.; Zhang, W.Y.; Ma, F.; Cheng, L.F.; Shi, F. Global distribution and development status of shale oil resources. *Miner. Conserv. Util.* **2019**, *39*, 126–134. [CrossRef]
4. Yang, L.; Jin, Z. Global shale oil development and outlook. *China Pet. Explor.* **2019**, *24*, 553–559.
5. Liu, F.; Zhu, X.; Liang, J.; Chen, G.; Liu, T. Characteristics of deep-water gravity flow petrographic development and its reservoir properties in the Yanchang Formation, Ordos Basin. *Mar. Geol. Front.* **2020**, *36*, 46–55. [CrossRef]
6. EIA. *Technically Recoverable Shale Oil and Shale Gas Resources: An Assessment of 137 Shale Formations in 41 Countries Outside the United States*; US Department of Energy: Washington, DC, USA, 2013.
7. EIA. *Annual Energy Outlook 2017 with Projections to 2050*; US Energy Information Administration: Washington, DC, USA, 2017.
8. Ma, L.; Xiao, H.; Tao, J.; Su, Z. An intelligent lithology classification method based on gradient boosting decision tree algorithm. *Oil Gas Geol. Recovery* **2022**, *29*, 21–29. [CrossRef]
9. Xu, H.; Cheng, D.; Xu, Y.H.; Yao, K.; Qiu, F.; Wu, X.; Lin, P. Formation lithology identification based on trenchless slurry performance testing system with weakly supervised learning. *Geol. Sci. Technol. Bull.* **2021**, *40*, 293–301. [CrossRef]
10. Xu, H.; Yao, K.; Cheng, D.; Song, Q.; Ma, Z.; Zhu, X.; Wu, X.; Zhao, G.; Cai, X. Identification of formation lithology based on trenchless drilling inspection system and random forest. *Geol. Sci. Technol. Bull.* **2021**, *40*, 272–280. [CrossRef]
11. Gu, Y.F.; Zhang, D.Y.; Bao, Z.D. Identification of lithology in dense sandstone reservoirs using a hybrid model CRBM-PSO-XGBoost. *Pet. Nat. Gas Geol.* **2021**, *42*, 1210–1222.
12. Xu, Z.H.; Ma, W.; Lin, P.; Shi, H.; Liu, T.H.; Pan, D.D. Intelligent recognition of lithology based on rock image migration learning. *J. Appl. Basic Eng. Sci.* **2021**, *29*, 1075–1092. [CrossRef]
13. Li, X.; Fan, X.Y.; Wang, Z.F.; Li, Y.X.; Chen, K.G.; Ma, X.L. Logging lithology identification method research based on PSO-SVM: A case study of Paleozoic (Pz) reservoir in K oil field, South Turgay Basin, Kazakhstan. *Prog. Geophys.* **2022**, *37*, 617–626. [CrossRef]
14. Gu, Y.F.; Zhang, D.Y.; Bao, Z.D.; Guo, H.X.; Zhou, L.M.; Ren, J.H. Lithology identification of dense sandstone formations using GS-LightGBM machine learning model. *Geol. Sci. Technol. Bull.* **2021**, *40*, 224–234. [CrossRef]
15. Ye, T.; Niu, C.; Wang, Q.; Gao, K.; Sun, Z.; Chen, A. Identification of paleosubduction metamorphic lithologies by the "composition-structure" classification method: An example from the Bohai Sea. *Rocky Reserv.* **2021**, *33*, 156–164.
16. Zhou, H.; Zhang, C.; Zhang, X.; Wu, Z.; Ma, Q. A capsule network-based lithology identification method for carbonate reservoirs. *Nat. Gas Geosci.* **2021**, *32*, 685–694.
17. Wang, Z. *Research on Lithology Recognition Method Based on Convolutional Recurrent Neural Network and Integrated Learning*; Yanshan University: Qinhuangdao, China, 2021. [CrossRef]
18. Gu, Y.F.; Zhang, D.Y.; Bao, Z.D. PSO-GBDT identification of lithology in dense sandstone reservoirs: An example of the long 4+5 section in the western Ji Plateau oilfield. *Miner. Rock Geochem. Bull.* **2021**, *40*, 624–634. [CrossRef]

19. Zhao, J.; Wang, W.M.; Yang, Z.D.; Zhang, M.L. Lithology identification method of volcanic rocks based on DBN model: An case in Chepaizi area. *Prog. Geophys.* **2022**, *37*, 328–337. [CrossRef]
20. Wu, Z.Y.; Zhang, X.; Zhang, C.H.L.; Wang, H.Y. LSTM recurrent neural network-based lithology identification method. *Rocky Oil Gas Reserv.* **2021**, *33*, 120–128.
21. Yuan, B.; Weijie, C.; Jin, Z.; Lijuan, L.; Feng, L.; Yongchang, G.; Baifa, Z. Addition of alkaline solutions and fibers for the reinforcement of kaolinite-containing granite residual soil. *Appl. Clay Sci.* **2022**, *228*, 106644. [CrossRef]